DOKTORI DISSZERTÁCIÓ
TÉZISEI

Fejlődő tudáshálózatok központjainak előrejelzése

SZÁNTÓ-VÁRNAGY ÁDÁM

Fizika Doktori Iskola
Doktori Iskola vezető: Prof. Tél Tamás

Statisztikus Fizika, Biológiai Fizika,
és Kvantumrendszerek Fizikája Program
Programvezető: Prof. Kürti Jenő

Témavezető: Farkas Illés J. D.Sc.

Biológiai Fizika Tanszék
Eötvös Loránd Tudományegyetem
Budapest
2019

Bevezető

Napjainkban az információs robbanás következtében nagyméretű adathalmazok gyűltek össze, emiatt különösen nagy szükség van az olyan eljárásokra, amelyek hatékonyan képesek feldolgozni azokat. Ennek a közös célnak az elméleti alapjait különböző tudományterületek képviselői fejlesztették ki. A statisztikus fizika a komplex rendszerek vislelkedésének mérése által járul hozzá ehhez, sajátos megközelítésével és tapasztalataival.

Jelen munkában négy gyakori kérdést vizsgálunk ezen a területen, gyakran többféle mérték együttes alkalmazása által. A közös cél mindezekben az, hogy behatóbb megértést nyerjünk ezekről a nagy adatsorokról, megtaláljuk a kiemelkedő fontosságú elemeiket ("központ"), és vizuális eszközökkel nyomon követhessük őket, ami – a tapasztalatok szerint – az emberi szemlélők számára a legalkalmasabb megoldás.

Különlegessége a bemutatott kísérleteknek, hogy teljesen általános bemeneti adathalmazokon képesek működni, ami lehetővé teszi, hogy bármilyen más körülmények között is rugalmasan alkalmazhassuk őket. Az eljárások alapjai a témák, vagyis azok a szavak, amik a publikációkban, cikkekben gyakran fordulnak elő, valamint ezek időbeli fejlődése. Gyakran az elemzést a hálózati struktúra vizsgálata egészíti ki.

Egy másik fontos eleme jelen kutatásnak a gyakorlatiassága, és az a tény, hogy nagy bemeneteken is hatékonyan működnek. Nem minden gyakori és jól ismert algoritmus rendelkezik ezzel a fajta skálázhatósággal. A bemutatott négy metódus közül az első kettőt, éppen emiatt, csak a legkisebb adathalmazon értékeltük ki (ami szintén viszonylag nagy, sok más adatsorhoz képest, mivel 400,000 rekordot és 4,7 millió kapcsolatot tartalmaz). Mind a négy kutatás esetében, az eljárások forráskódja nyilvánosan elérhető (a `http://topinav.elte.hu/` címen).

Eredmények

1 Téma diagramok / idősorok

1.1 Témák kódjai: számsorozatok gyártása diagramok alapján

Bevezettem a téma kódok fogalmát, ami az idősorok jól ismert koncepciójának leegyszerűsítése, két egyszerű paraméter segítségével (ami az eredeti diagram *magasságának* és *szélességének* felosztásának feleltethető meg).

A *téma kód* egy rövid számsorozat, ami egy teljes téma diagramot / idősort ír le. Egyszerűségének köszönhetően különösen hatékonyan segít praktikus alkalmazásokat, például hasonló idősor keresését (ld. lent, 4.3), vagy idődiagramok osztályzását és jóslását (ld. lent, 5.1).

2 Egymás idézettségét segítő (felpörgető) cikkek

*A felpörgetés (*boost*) azt jelenti, hogy az idézett cikk hivatkozásszáma megnő, az egyik idéző cikknek köszönhetően. Ennek a kutatásnak a motivációja az, hogy minőségileg kiemelkedő publikációk halmazát segítsen beazonosítani.*

2.1 Felpörgetés mérőszámai

Találtam cikkeknek egy párját, ahol egyik a másikat felpörgeti, majd három mérőszámot vezettem be, ami hasonló párok keresésére alkalmas.

Megfigyeltem, hogy nem elegendő, hogy az idézett és idéző cikk közös hivatkozásai jelentős hányadot képezzenek, hanem ezenkívül az is szükséges, hogy a két publikáció megjelenése között nagy idő teljen el, valamint az idézett cikknek önmagában is magas legyen a hivatkozottsága. E két megszorítás bevezetése által hatékonyan eltávolítot-

tam azokat a találatokat, amik csak látszólagos példái a felpörgetés jelenségének. Az American Physical Society publikációin futtattam az eljárásokat, melyből nem adódott olyan pár, ami megközelítené a felpörgetésre talált első példa cikkpárat (ami nem ebből az adathalmazból származik).

2.2 Felpörgetési hálózat elemzése

Elemeztem az összes olyan cikkpár által alkotott hálózatot, melyek valóságos felpörgetést mutattak a fent bevezetett mérőszámok alapján. Ebből nyolc jelentősebb részhálózat adódott, melyek mérete 9 és 21 publikáció között mozog.

Ezen elemzés céljából küszöbértékeket állítottam be, melyhez az adatsor vizsgálata során használt statisztikákat használtam fel. A legtöbb talált komponens rámutat némely központi fontosságú publikáció jelentőségére, melyek szemmel láthatóan kiemelkedő szerepet töltenek be a többi publikáció között. Ezeken a komponenseken kívül a hálózat kisebb, triviális részeket tartalmaz.

3 Egész témákat felpörgető cikkek

3.1 Meredekség-visszaesés beazonosítása

Minden témakör diagramját elemeztem olyan mérőszámok segítségével, amik beazonosítják egy téma tartós növekedését, az említéseinek száma alapján.

A *meredekség* mérésének célja, hogy beazonosítsa egy adott téma iránt rövid idő alatt megnövekvő érdeklődést, míg a *visszaesés* célja, hogy megbizonyosodjon róla, hogy ez a megnövekedett érdeklődés valóban tartós maradt-e, és nemcsak a véletlen zaj végeredménye. Azokban az esetekben, amikor a növekedés a vizsgált időtartam végén történt, a visszaesés a diagram tükrözése által vizsgálható ("inverz visszaesés"). Az alacsony visszaesés és magas meredekségi értékek tényleges ugrást jeleznek. A két mérték együttes vizsgálata közös ábrán való megjelenítés által történik.

3.2 Perkoláció és lefedettség a téma alhálózatában

Minden ugrást mutató témához elemeztem azon cikkek hálózatát, melyek az adott időszakban a téma kulcsszavát címeikben tartalmazzák.

Cikkek olyan kis halmazát kerestem, melyek képesek magyarázni nagy változást a téma forgalmában. Ehhez megkerestem az ezen cikkekből adódó hálózat legnagyobb komponensét, és azon cikkek halmazát, melyek minden más, komponensben lévő cikkre, közvetlenül vagy közvetetten, hivatkoznak. Megmutattam, hogy ezen mindkét mérték pozitív korrelációt mutat a téma meredekségével (ld. fent).

4 Hasonlósági mértékek összehasonlítása

4.1 Hasonlósági mértékek általános kiértékelése

Olyan eljárást javasoltam, amely tetszőleges hasonlósági mértékek összehasonlítására alkalmas, a legközelebbi szomszédtól mért távolság alapján. Illusztráltam az algoritmus működését 3 egyszerű mérőszám segítségével.

Két témakör hasonlósága mérhető 1. szövegi hasonlóság alapján (ezen belül kétféle változat áll rendelkezésre), 2. hálózati struktúra alapján, vagy 3. a téma időbeli népszerűsége alapján. Egy hasonlósági mértéket sikeressége mérhető azon szavak száma által, melyekben az adott mérték közelebbi szomszédot talált a többi mértéknél. Ehhez hasonló általánosságú definíció, melyhez további bemeneti értékeket nem kell megadni, nem található a meglévő szakirodalomban.

4.2 Hatványfüggvény alapú normálás

Különböző hasonlósági mértékeket hasonlítottam össze, melyek különböző skálákon vesznek fel értékeket. Egy megfelelően megválasztott normálás, mely hatványfüggvényre való illesztésen alapszik, alkalmas arra, hogy mindet közös skálán jelenítse meg.

A két szövegalapú és a hálózatalapú hasonlóság tipikusan nagy egész értékeket vesz fel. Mivel ezek többségében hatványfüggvény-jellegű eloszlást követnek, ami log-log skálán lineáris függvényként jelenik meg, ezért az elemek elhelyezkedése az így adódó lineáris skálán egy természetes választás a hasonlósági értékek 0 és 1 közé való skálázáshoz.

4.3 Hatékony legközelebbi szomszéd-keresés dobozok segítségével

Kifejlesztettem egy módszert, ami megtalálja a leghasonlóbb idődiagramot a jelenleg kiválasztotthoz képest. A felhasznált paraméterek az adatsor alapján vannak optimalizálva, melynek köszönhetően a módszer alkalmas nagy számú diagramon való futtatásra, anélkül, hogy minden diagramot mindegyikhez kellene hasonlítani.

Ez a feladat a tipikus példája egy könnyen érthető, látszólag kézenfekvő problémának, melynek gyakorlati megoldása kreatív gondolkodást, feladatmegoldó képességet igényel. Enélkül az algoritmus hamar négyzetes időigényűvé válik, melynek alkalmazása szinte lehetetlen bármilyen nagyobb adatsor esetén. Ehhez az 1.1-es pontban bevezetett téma kódokat használtam, kétféle paraméter választással: az egyik a páronkénti hasonlóság számításához szükséges, a másik ahhoz, hogy a szavakat "dobozokhoz" rendelje, melyeket utána az algoritmus összehasonlít minden szomszédos dobozzal.

5 Idősor predikció legközelebbi szomszédok alapján

Kifejlesztettem egy idősor predikciós eljárást, melynek alapja az, hogy a diagramokat növő, csökkenő és stagnáló csoportokba osztályozza. Az alapötlet az, hogy olyankor számítunk egy szó gyakoriságának növekedésére a következő pl. 10 évben, ha szomszédainak átlaga is növekedett az előző 10 évben.

A vizsgált időtartam két, egyenlő részre oszlik: múlt és jövő. Egy egyszerű osztályozó határozza meg, hogy az adott diagram nő, csökken vagy stagnál. Ez a kiértékelés történik a vizsgált szó múltjára, szomszédainak átlagára, és e kettő közötti különbségre.

Minden lehetséges bemeneti kombinációhoz a leggyakoribb kimenetet választjuk, mint a jóslás értékét (ez *1R-osztályozó* néven is ismert). Az így kapott eljárás arra is alkalmas, hogy váratlan eredményeket jósoljon, például amikor a szó múltbeli csökkenése korrelációt mutat a szó jövőbeli növekedésével. A többféle bemeneti változó hatékonynak bizonyult a jóslás bizonyosságának szignifikáns növelésére, azokban az esetekben, amikor az egyszerű bemenet önmagában nem ért el kellően magas találati értékeket.

5.1 Szomszédok reprezentáns halmazának kiválasztása

Kifejlesztettem egy eljárást arra, hogy hogyan állapítsuk meg adatsoronként a legközelebbi szomszédok számát, melyet a predikcióhoz figyelembe veszünk. Ezál-

tal jelentős számításigény takarítható meg, és nem szükséges minden szó minden szomszédjának adatait figyelembe venni. Meglepő módon, amint az ábrák alapján látható, minden adatsorhoz van egy egyértelmű érték, ahol a szükséges szomszédok számának határa meghúzható.

Az összes szóból választott, kellően nagy méretű minta vétele által, teljes mértékben kiértékeltem a kiválasztott szavak összes szomszédját. Ezen szomszédokat az eredeti szótól vett távolság szerinti sorrendbe rendeztem, és minden N értékhez az első N szomszéd átlagos diagramját számítottam ki. A különbség az így kapott diagram és a végleges diagram között gyorsan konvergált 0-hoz (N függvényében), és miután minden szóra ilyen módon megállapítottam a szükséges küszöbszámot, megjelenítettem az összes szóra vonatkozó eloszlást.

5.2 Bizonytalanság figyelembe vétele a jóslás végeredményében

Kiegészítettem a jóslás során használt egyszerű döntési szabályok definícióját úgy, hogy a meglévő kimeneti értékek következő logikai kompozícióit is tartalmazzák: *"növekedés VAGY stagnálás"*, ill. *"csökkenés VAGY stagnálás"*.

Ez a kompromisszum szükséges volt az eredmények fényében, ugyanakkor alkalmassá teszi az eljárást arra, hogy olyankor is mondjon bizonyos korlátozott predikciót, amikor a pontos végeredmény nem jósolható meg. A bemeneti értékek alapján tudható az, hogy milyen arányban adódik növekvő, csökkenő és stagnáló kimenet. Ha e három arányérték közül kettő közel esik egymáshoz, akkor bizonytalan szituációról beszélünk. Kiértékeltem különféle lehetséges paraméterértékeket a *nullatűréshez*, melynek célja, hogy optimalizálja, hogy mennyire közel kell két értéknek esnie egymáshoz, ahhoz, hogy bizonytalannak definiáljuk a szituációt.

Publikációk

A disszertáció témájához közvetlenül kapcsolódó cikkek

[1] Szántó-Várnagy Ádám, Farkas Illés J. "Forecasting turning trends in knowledge networks". In: *Physica A: Statistical Mechanics and its Applications* 507 (2018), pp. 110–122. ISSN: 0378-4371. DOI: https://doi.org/10.1016/j.physa/2018.05.055. URL: http://www.sciencedirect.com/science/article/pii/S0378437118305995.

[2] Szántó-Várnagy Ádám, Pollner Péter, Vicsek Tamás, Farkas Illés J. "Scientometrics: untangling the topics". In: *National Science Review* 1.3 (2014), p. 343. DOI: http://dx.doi.org/10.1093/nsr/nwu027. URL: https://academic.oup.com/nsr/article-lookup/doi/10.1093/nsr/nwu027.

További publikációk

[3] Szántó-Várnagy Ádám, Pollner Péter és Farkas Illés J. "Measuring originality in knowledge networks". In: *Lecture Notes in Computer Science* 9197, Paper: 156799 (2015).

[4] Farkas Illés J., Szántó-Várnagy Ádám, és Korcsmáros Tamás. "Linking Proteins to Signaling Pathways for Experiment Design and Evaluation". In: *PLOS ONE* 7.4 (Apr. 2012), pp. 1–5. DOI: http://dx.doi.org/10.1371/journal.pone.0036202. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0036202.